WIDER TECHNICAL NOTE | 17/2021

SOUTHMOD – simulating tax and benefit policies for development

Iterative proportional fitting for reweighting input data in SOUTHMOD microsimulation models

David McLennan*

September 2021

wider.unu.edu

Abstract: This note describes how the counterfactual datasets in Lastunen et al. (2021) have been reweighted for each country of the analysis to the 'pre-crisis' time point of 31 March 2020. The procedure consists of five main steps. The note forms part of a series of technical notes that complement Lastunen et al. (2021).

Key words: COVID-19, SOUTHMOD

JEL classification: C80

Acknowledgements: The work presented is based on SOUTHMOD microsimulation models, developed, maintained, and managed by UNU-WIDER in collaboration with the EUROMOD team at the Centre for Microsimulation and Policy Analysis (University of Essex), SASPRI (Southern African Social Policy Research Insights), and local partners in selected developing countries (Ethiopia, Ghana, Mozambique, Tanzania, Uganda, Zambia, Ecuador, and Viet Nam) in the scope of the SOUTHMOD project. This technical note is an extension of a SOUTHMOD research project on the impact of COVID-19 in Africa (Lastunen et al. 2021) and Viet Nam.

Related publications: at the end of the paper

---

* Southern African Social Policy Research Insights (SASPRI), Hove, UK; david.mclennan@saspri.org

This note has been prepared within the UNU-WIDER project SOUTHMOD – simulating tax and benefit policies for development Phase 2, which is part of the Domestic Revenue Mobilization programme. The programme is financed through specific contributions by the Norwegian Agency for Development Cooperation (Norad).

Information and requests: publications@wider.unu.edu

https://doi.org/10.35188/UNU-WIDER/WTN/2021-17

United Nations University World Institute for Development Economics Research

Katajanokanlaituri 6 B, 00160 Helsinki, Finland

The United Nations University World Institute for Development Economics Research provides economic analysis and policy advice with the aim of promoting sustainable and equitable development. The Institute began operations in 1985 in Helsinki, Finland, as the first research and training centre of the United Nations University. Today it is a unique blend of think tank, research institute, and UN agency—providing a range of services from policy advice to governments as well as freely available original research.

The Institute is funded through income from an endowment fund with additional contributions to its work programme from Finland, Sweden, and the United Kingdom as well as earmarked contributions for specific projects from a variety of donors.

1 Objectives

The following short note forms part of a series of technical notes that complement Lastunen et al. (2021). The preparation of pre-crisis, counterfactual datasets in Lastunen et al. (2021) requires reweighting the newest input dataset from each country to the 'pre-crisis' time point of 31 March 2020.

The reweighting procedure recalibrates the survey weights in each of the input datasets so that the weighted population totals correspond to the estimated population totals and demographic profiles. It also holds constant the labour market profiles as enumerated in the input data.

The reweighting process we implemented uses the technique of iterative proportional fitting (raking). The Stata .ado file 'ipfraking' was utilized for this purpose. The procedure consisted of five main steps, discussed in Section 2.

2 Analysis steps

2.1 Preparing the external population control totals relating to demographic profile

This step consists of extracting the relevant population estimates from the official statistics provided by country team partners. If possible, population estimates were extracted for the year of survey enumeration and for mid-2019 and -mid-2020. For Zambia, Tanzania, and Uganda, population estimates by quinary age and sex were available from the national statistical agencies for the respective year of survey enumeration as well as both mid-2019 and mid-2020. For Mozambique, population estimates were available for mid-2019 and mid-2020, but not for the year of survey enumeration (2015). A broader range of population estimates were available for Zambia than for the other countries considered. For instance, the Zambian estimates included disaggregation by province, district, and rural/urban location. As the three simulated benefits in Zambia contain elements of geographical targeting, these additional demographic profiles were prepared alongside the standard quinary age/sex profiles for Zambia. In order to derive population estimates for the agreed 'pre-COVID' time point of 31 March 2020, it was necessary to interpolate between the population estimates for mid-2019 and mid-2020. A simple linear interpolation approach was used to derive the estimates for 31 March 2020.

2.2 Preparing the microsimulation input datasets

This step consists of producing derived variables in the microsimulation input dataset to match the categories of the external control totals. A new variable 'agesex' was derived in each of the input datasets to categorize the survey respondents according to quinary age group and sex. For Zambia, a broader range of categorical variables was derived in order to reflect the broader range of population control totals. To enable the labour market profile to be explicitly controlled during the reweighting process, a new composite variable was created by combining the 'current economic status' variable ('les') with the 'primary occupation code' variable ('loc'). This new 'les_loc' variable disaggregated the 'self-employed' and 'employee' economic status groups according to their reported occupational type. Although slight variations between country input datasets were necessary to reflect country-specific coding schemes, the general categorization applied across countries included farmers, 10 occupation categories for the self-employed, 10 occupation categories for employees, pensioners, those unemployed, students, home makers, those inactive, others, and those aged under five years. For Zambia and Uganda, the final category relates to children aged 0–4 years. In Mozambique and Tanzania, slight differences in the coding scheme mean that the final category relates to children aged 0–14.

2.3 Preparing the external population control totals relating to labour market profile

Weighted population shares by labour market category ('les_loc') were calculated for each country input dataset and these population shares were then applied to the population totals derived from the population estimate external control totals produced in Step 1 of the reweighting procedure. Scaling the labour market profile to the respective population totals ensures that the sum of the labour market categories matches the sum of the age/sex categories, thereby maintaining internal consistency between the two sets of external controls. Although this is not strictly necessary for the reweighting process, it provides methodological clarity as the adjustments to the control totals are explicitly defined rather than being implicitly generated during the reweighting procedure.

2.4 Running the iterative proportional fitting procedure

The 'ipfraking' Stata .ado file was used to operationalize the iterative proportional fitting procedure. For those countries with population estimates available for the year of survey enumeration as well as for mid-2019/mid-2020 (i.e. models for Zambia, Tanzania, and Uganda), the first round of iterative proportion fitting was configured to reweight the input dataset so that the demographic profile of the reweighted survey counts matched the demographic profile in the external population control total(s) for the year of survey enumeration, holding constant the labour market profile. The process commences with the importation of the external control totals (demographic profile and labour market profile) from the specified spreadsheets (Step 1) into Stata matrices.

The 'ipfraking' command is then configured according to the country-specific specification of the input dataset. Running a 'summerize' command in Stata on the original survey weight variable (dwt) in the input dataset reveals the range and distribution of the original weights. Next, this information is used to inform the setting of the 'trimming' parameters in the 'ipfraking' command, which allows for the setting of the absolute and relative trim limits within which the 'ipfraking' procedure is forced to operate.[1,2] The second part of the 'ipfraking' procedure entailed reweighting the survey to the demographic and labour market profiles for 31 March 2020. This time, for Zambia, Tanzania, and Uganda, a 'summerize' command was run on the rebased survey weights calculated as described above, and these weight distributions informed the configuration of the trimming parameters in the rebasing to 31 March 2020. For Mozambique, the original survey weights were assessed in the 'summerize' command, and these informed the specification of the trimming parameters.

[1] There is no hard rule in terms of how the trimming parameters should be configured, so these parameter values were specified separately for each input dataset according to the distribution of original survey weights and to ensure convergence of the 'ipfraking' procedure. The setting of trimming parameters inevitably involved an element of 'trial and error', with the parameter values gradually relaxed or tightened and the results from the 'ipfraking' procedure reviewed. The objective was to achieve convergence of the procedure within the specified tolerances, whilst minimizing the magnitude of changes to individual weights and minimizing changes to the overall range and distribution of the weights.

[2] For the Mozambique model, no reliable population estimates were available from the national statistical agency for the year of survey enumeration, so this initial 'ipfraking' process was not possible.

2.5 Performing quality assurance tests on the reweighted input datasets

The final part of the 'ipfraking' procedure entailed, first, reviewing the outputs from the reweighting exercise to ensure that the weighted counts by demographic profile and labour market profile corresponded to the external control totals and, second, reviewing the simulated outputs from the models when underpinned by the reweighted input datasets to check plausibility of the simulated revenues, expenditure, poverty rates, and inequality measures. A further labour market check was undertaken to ascertain the share of weighted survey respondents classified as being in 'formal' or 'informal' employment.

Although the formal/informal status was not considered in the 'ipfraking' procedure, checking the output across this variable provides an additional means to verify that the results of the reweighting are plausible. The validation work indicates that the re-based survey weights for the agreed 'pre-COVID' time point of 31 March 2020 are in line with each of the external control totals and the distributions are plausible when considered in the context of the distributions of the original survey weights in the input datasets. The preliminary validation results from the four countries are available upon request. Based upon these conclusions, the re-based survey weights can now be merged into the original input datasets and these datasets can be used as a basis for applying the 'COVID-induced' labour market adjustments.

Reference

Lastunen, J., P. Rattenhuber, K. Adu-Ababio, K.H. Gasior, H.X. Jara, M. Jouste, D. McLennan, E. Nichelatti, R.C. Oliveira, J. Pirttilä, M. Richiardi, and G. Wright (2021). 'To the Rescue? The Mitigating Role of Tax and Benefit Rescue Packages for Poverty and Inequality in Africa amid the COVID-19 Pandemic'. WIDER Working Paper 2021/148. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2021/088-7

Related publications

Lastunen, J. et al. (2021). 'To the Rescue? The Mitigating Role of Tax and Benefit Rescue Packages for Poverty and Inequality in Africa amid the COVID-19 Pandemic'. WIDER Working Paper 2021/148. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/2021/088-7

Lastunen, J. (2021). 'Deriving Shocks to Household Consumption Expenditures from the Associated Income Shocks Resulting from COVID-19.' WIDER Technical Note 2021/16. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/WTN/2021-16

Gasior, K. et al. (2021). 'Full-Year Adjustment for Modelling COVID-19 Policies in SOUTHMOD Tax-Benefit Microsimulation Models'. WIDER Technical Note 2021/18. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/WTN/2021-18

Oliveira, R.C. et al. (2021). 'Imputation Methods for Adjusting SOUTHMOD Input Data to income Losses Due to the COVID-19 Crisis'. WIDER Technical Note 2021/19. Helsinki: UNU-WIDER. https://doi.org/10.35188/UNU-WIDER/WTN/2021-19